Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 19

5-1-2004

Validation Studies: Matters Of Dimensionality, Accuracy, And Parsimony With Predictive Discriminant Analysis And Factor Analysis

David A. Walker
Northern Illinois University, dawalker@niu.edu

DOI: 10.22237/jmasm/1083370740

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 186-199

1538 – 9472/04/$95.00

Validation Studies: Matters Of Dimensionality, Accuracy, And Parsimony With Predictive Discriminant Analysis And Factor Analysis

David A. Walker
Educational Research and Assessment Department
Northern Illinois University

Two studies were used as examples that examined issues of dimensionality, accuracy, and parsimony in educational research via the use of predictive discriminant analysis and factor analysis. Using a two-group problem, study 1 looked at how accurately group membership could be predicted from subjects' test scores. Study 2 looked at the dimensionality structure of an instrument and if it developed constructs that would measure theorized domains.

Key words: Predictive discriminant analysis, factor analysis, dimensionality, accuracy, parsimony

Introduction

The first study in this article has two intentions. First, if there is an interest in the degree to which group membership, based upon a set of predictor variables, can be predicted the question posed may be: How accurately can group membership in either Average grade point average (GPA) or Above Average GPA from the subjects' Florida College Level Academic Skills Test (CLAST) scores? A second question may be: In terms of their relative contribution to classification accuracy, how well can a ranking of the predictor variables predict if a subject taking the CLAST is going to be in the Average GPA group or the Above Average GPA group?

Study 1.

The CLAST is an achievement test that was first implemented by the Florida State Board of Education (SBE) in 1984 as part of its educational accountability measures. The test is comprised of four subtests in mathematics, reading, writing, and essay that purport to measure students' academic proficiency, by the completion of the sophomore year, in the areas of computation and communication. The CLAST is administered three times a year in October, February, and June. Students who have accrued a minimum of 18 semester hours may apply to sit for the test. Institutions may require students to pass 3 subtests before they can earn more than 60 degree credits and/or pass all 4 subtests before obtaining 96 degree credits toward a baccalaureate degree.

Subtests, however, can be taken as many times as needed until passed. To receive an associate in arts degree from any of Florida's 28 public community colleges or obtain admission to upper-division status in any of Florida's 11 public, 4-year institutions, a student must pass all subtests of the CLAST or receive one of many exemption options (Florida Atlantic University, 2002; Florida Department of Education, 2000).

Exemptions from any of the three communication subtests are predicated on attaining a 2.50 GPA in two designated college-level English courses. Exemption from the mathematics portion is based on a 2.50 GPA in two defined courses. Also, an ACT score of 21 in mathematics, a 22 in reading, a 21 in English, or an SAT score of 500 in quantitative and/or verbal are approved exemptions. A documented learning disability or physiological impairment, or if a student has already earned a Bachelor's

David Walker is an Assistant Professor of Educational Research and Assessment at Northern Illinois University. His research interests include structural equation modeling, effect sizes, factor analyses, predictive discriminant analysis, predictive validity, weighting, and bootstrapping. Email: dawalker@niu.edu.

degree and is seeking a second undergraduate degree, will merit an exemption (Florida International University, 2002; University of South Florida, 2002). It should be noted that such exemptions have the ability to reduce the internal and external score validity of the CLAST.

The subtests measure students' academic proficiency in lower-division course work in the general areas of mathematics, reading, writing, and essay. The mathematics subtest includes selection-type items (i.e., multiple-choice) in the following areas: algebra, arithmetic, geometry, logical reasoning, measurement, probability, and statistics. The reading subtest has multiple-choice items that measure two areas: literal comprehension and critical comprehension. The English portion of the CLAST also uses multiple-choice items and measures students' skill levels in the areas of word choice, sentence structure, grammar, spelling, capitalization, and punctuation. Scores for the mathematics, reading, and English subtests range from 200 to 400 points.

The SBE has changed the cut scores for passing these 3 subtests from a minimum score of 260 in 1984 to a present score of 295. Current mean averages for first-time examinees from the 1999-00 academic year show that mathematics had a 3 administration average of 299, reading was 305, and English was 309 (Florida Atlantic University, 2002; Florida Department of Education, 2000).

The essay test allows students to choose from two topics and write about one of these. Essay writing measures students' skills in the areas of composition, effective language use, and the dissemination of ideas. Using a holistic rubric, two trained readers rate each essay test. Essay scores range from 2 to 12 points. In 1984, the original cut score was a 4, however; the current minimal score has been changed to a 6. From academic year 1999-00, the mean average for the essay test was a 7 (Florida Department of Education, 2000; Indian River Community College, 2002).

Methodology

The four predictor variables were the subtests on the CLAST: mathematics, reading, English, and essay. The criterion variable was undergraduate GPA, where 4.00 = A, 3.00 = B, 2.00 = C. There were no GPAs below 2.00 because to be in the sample as a recent graduate of a Florida four-year public institution, a participant needed at least a 2.00 to graduate. Thus, GPA was operationalized as a comprehensive academic performance measure of students' cognitive abilities in their entire degree program of study. GPA has been used as a criterion variable and is often influenced by many factors such as the facility or difficulty level of course content, student effort, instructor competency, and student involvement, or not, in co-curricular activities. More considerably, GPA is a variable that has been cited as a measure of students' cognitive abilities, especially in the areas of verbal and quantitative skills (Brown & Campion, 1994; Roth & Bobko, 2000; Wolfe & Johnson, 1995).

Reliability

Using the Kuder-Richardson 20 method, the reliability of the CLAST subtest scores for the 3 administrations in 1999-00, along with standard error of measurements shown in parentheses, were .83 (3.03), .84 (3.02), and .86 (3.07) for mathematics; .74 (2.74), .83 (2.38), and .77 (2.37) for reading; and .71 (.2.21), .67 (2.17), and .68 (2.21) for English. The essay subtest score reliability, pertaining to the trained readers' ratings of each of the two essay topics, was measured through inter-rater reliability (IRR) derived from a six-point holistic scoring rubric. For the 3 administrations in 1999-00, the IRR scores for the 2 essay topics were .86, .85, and .86 for topic 1 and .86, .87, and .83 for topic 2 (Florida Department of Education, 2000).

Results

Using a resampling cross-validation technique, the Leave-One-Out (L-O-O) rule or U method (Huberty, 1994; Lachenbruch & Mickey, 1968), the subset of all possible variables were analyzed for the purpose of parsimony, theoretically where "simpler hypotheses are more falsifiable," (Meehl, 1993, p. 5) and to increase the cross-validation accuracy of the proposed model (Lieberman & Morris, 2004; Morris & Meshbane, 1995). Morris and

Meshbane's FORTRAN program (Huberty, 1994, Morris & Meshbane, 1995) for an all subset analysis to yield the best L-O-O hit rate for predictor selection, or 2^p -1 where p are the predictors, was conducted.

Of the initial four variables considered, two predictors were deleted that did not contribute to high predictive accuracy (i.e, math and reading). Thus, only writing and essay were retained as components of a parsimonious and more credible model (i.e., in terms of the population). That is, there were 4 predictor variables for the 2-group problem, which meant that there were 15 all possible subset analyses (i.e., 2^4 -1). When the number of predictors in the best subset of 2^p -1 emerged, the maximum hit rate increased by almost 1.00% to 58.40% from the second best hit rate of 57.47% with 3 predictors (i.e., writing, essay, and math), and, thus, parsimony with increased accuracy was achieved. Other variations within the all possible subset analyses yielded a range of maximum hit rates between 52.80% and 58.40%.

With the L-O-O method, it has been noted that a minimum sample size can be calculated as N = 3kp or a large sample size of N = 5kp, where k is the number of groups and p is the number of predictors, and the 3 or 5 derived from the n/p ratio (Huberty, Wisenbaker, & Smith, 1987). The study's sample size of 750 subjects was adequate. Multivariate normality of the data and equality of covariance matrices of the groups were met, with a normal-based rule establishing normality via a review of normal probability plots for data in each of the two groups (Huberty & Lowman, 1998).

A significant degree of discrimination separating the two groups of study was confirmed. As a classification rule, equal prior probabilities external to the sample were established at .50 (q_1) /.50 (q_2), which measured the probability of population membership in either group and equal cost of misclassification for the two populations. The choice of equal priors assumed that the accuracy of this decision was based on estimated priors from the population and not the sample. It has been noted that adjusting for unequal priors based on an estimation from the group size of the sample can be misleading and potentially costly in terms of decreased model classification accuracy (Meshbane & Morris, 1996).

The GPA for subjects classified as Average ranged between a "C" (i.e., 2.00) and "B-" (i.e., 2.99), and the GPA for subjects classified as Above Average ranged between a "B" (i.e., 3.00) and "A" (i.e., 4.00). The cut point chosen for the two groups was the median GPA for all of the subjects in the study at 3.00. Thus, those subjects with GPAs below this cut point were grouped as Average and coded as a 0, and those with GPAs equal to or above this cut were grouped as Above Average and coded as a 1 (cf. Press & Wilson, 1978).

Table 1. Predictive Discriminant Analysis: Linear External Classification.

Cross-Validation L-O-O

	Average GPA	Above Average GPA	Total
Average GPA	168 (62.92%)	99 (37.08%)	267
Above Average GPA	213 (44.10%)	270 (55.90%)	483

58.40 of cross-validated grouped cases correctly classified.

The results from Table 1 present the L-O-O rule that was established as a bias correction method for classification error rates. L-O-O took 1 subject out of the sample and developed a rule on the other 749 subjects and then took another subject out and developed a rule on the other 749, and so on. This linear, external classification rule was applied to all subjects in the sample so that rules were built on all 750 (Huberty, 1994; Lachenbruch, 1967). From an SPSS (v. 12.0) analysis, table 1 presents the accuracy of the model on cross-validation, meaning how well does this model apply to subjects from the population or its generalizability.

For Average GPA, there were 168 or 62.92% subjects (90% CI for a Binomial Parameter = .578, .678; SE = .03) classified as Average or hits and 99 or 37.08% (CI = .322,

.422) that were predicted as Above Average or misses. For the Above Average GPA group, there were 213 or 44.10% subjects (CI = .403, .479; SE = .02) misclassified as Average or misses and 270 or 55.90% (CI = .521, .597) that were predicted as Above Average or hits. In terms of total precision for all of the subjects, there was 58.40% accuracy (CI = .554, .614; SE = .02). The model correctly classified a little over half of the cases, with a total group error rate estimate of 41.60% (CI = .386, .446).

When assessing each variable's contribution to the discriminant function, the standardized canonical discriminant function coefficients (weights) indicated that writing's relative importance in predicting GPA was .716 followed by essay at .514. Predictor importance was also noted via another method when writing, for example, was taken out of the model, which produced the lowest hit rate for total group accuracy at 52.80% (cf. Huberty & Lowman, 1998). The order of the response variables' contribution toward predictive accuracy indicated how the predictor variables should be arranged. In terms of structure coefficients, the largest absolute correlation associated with the discriminant function was writing at .872, with essay at .731.

In regard to particular cases that may be fence riders, or subjects that were classified correctly, but when their probabilities were reviewed, confidence waned in terms of proper classification, the probability split between highest group and second highest group was established at .52/.48. Of the 750 subjects, 32, or 4.27%, were deemed fence riders. Outliers were determined to be cases that had typicality probabilities less than .10. That is, although a subject was classified correctly with confidence, it appeared to be atypical of that group and hence garnered a low probability. Of the 750 subjects, 35, or 4.67%, were estimated to be outliers. The fence riders and the outliers were kept in the data and analyzed because omitting them may have inflated the hit rate of the model, which potentially could have yielded a model that was more accurate than in actuality.

Using a proportional chance criterion, Huberty's (1994) Z statistic was calculated from a FORTRAN program (J. D. Morris, personal communication, March 13, 2003) to determine if expected hit rates were exceeded.

$$Z = (o-e)/[e(n-e)/n]^{1/2} \quad (1)$$

o = observed frequency
e = expected frequency
n = number of subjects

This test is a one-tailed test because there is little interest in whether the hit rate was significantly below expectation. The null hypothesis was that the hit rate is what would be expected by chance (e.g., .50 x 267 + .50 x 483 = 375). The alternative hypothesis was that the present hit rate is better than chance expectance. With an observed hit rate of 438 (i.e., 168 + 270), the Z of 2.34 ($p < .02$) for the total sample occurred because this hit rate was above expectation, which offers some evidence that the null should be rejected or that classification by the discriminant function resulted in more hits than random assignment by prior probabilities.

However, when the Z value for each group was examined, a different inference emerged. The Z value for the Average GPA group was very large and statistically significant at 9.32 ($p < .001$), but the Z for the Above Average group was .00 and not statistically significant ($p > .05$). The reason this model appeared to be better than chance was that it was quite good at predicting the Average GPA group, but very poor at predicting the Above Average GPA group based on subjects' CLAST scores. That is, the percentage improvement over chance for the Average GPA group was 42.42% and for the Above Average GPA group was -23.87%. The percentage of improvement over chance for the total sample was only 9.27%. Thus, the classification of the two groups was only slightly better, by 9%, than would have been accomplished by chance.

To add to this argument from a different perspective, and also to address the issue of the intermediate inequality of group sizes, the model was looked at via a maximum chance criterion (max (q_1, q_2)) (Huberty, 1994). The maximum chance criterion assigned all of the subjects to the largest group for this study, the Average GPA group, as a criterion for a hit rate better than chance. The Z value was -.08, which meant that the model's hit rate was not better than chance. Further, the percent improvement over

chance for the total sample was -16.85%. Thus, this model did not have good accuracy for the two-group classification problem using either of the chance criteria proposed.

Huberty's (1994) effect size measure, the I statistic, was calculated to determine the

$$I = (1-e) - (1-o) / 1-e \quad (2)$$
$$= o-e / 1-e$$

percentage correctly classified exceeding chance. The Average group had an I = .258, the Above Average group had an I = .118, and the total model had an I = .168. Previous research (Huberty & Lowman, 2000) conducted on I indicated that these values should be regarded as having a low effect, except for the medium effect of the Average group, in terms of their ability to measure proportional reduction in error, meaning, for instance, that the total model had roughly 16% less misclassifications than would have occurred if just classified by chance.

Conclusion

The addition of many more exemptions on the CLAST created a problem where it was supposed that students from various colleges within a university could have opted out of the test, leaving the study with a more homogeneous sample (i.e., participants from only a few colleges who did not have as many exemption options).

For future institutional decisions related to students' academic success, the PDA model chosen for this study, which was parsimonious and contained two estimators of the CLAST subtest scores to classify students into one of two alternative populations consisting of Average GPA or Above Average GPA, was not accurate enough across all groups, or for each group, and its total sample hit rate was only slightly better than chance. Overall, the CLAST subtest scores did not estimate effectively academic success in terms of predicting GPA. In fact, the predictors' relative contribution ranged within a moderate level of ordered importance from writing (.716) to essay (.514), both of which were also rank-ordered as first and second most important using a variable deletion method, with 2 unimportant variables (i.e., math and reading) removed because classification accuracy did not diminish without their presence in the model. Thus, CLAST score use by institutions as a general measure of educational accountability, specifically in the instance as a mode to estimate high academic success, does not appear to be an effective model.

Study 2.

The New Teacher Academy (NTA) was created as a link to Florida's A+ Plan for K-12 public schools in Broward County, which during academic year 2001-02 enrolled 260,892 students (Broward County Public Schools, 2003) making it one of the 10 largest school districts in the United States. Specifically, the NTA was initiated to assist new teachers in Broward County Public Schools with bolstering their performance levels in the classroom as a measure of accountability, but also as a means of professional development in the sense of sustained, active development (Fullan, 2000; Garet, Porter, Desimone, Birman, & Yoon, 2001).

Further, to address the challenge of hiring more non-education major teachers to educate the increasing student enrollments within Broward's K-12 system, NTA was contrived to support these new teachers' development and overall preparedness in the classroom. In this manner, the NTA could be thought of as an approach for early professional development, but also as an agent for "teacher change" (Clarke & Hollingsworth, 2002).

A cross-functional planning committee, along with survey responses from educators in various capacities throughout the Broward County Public School system, assisted in identifying critical domains that all new teachers should know and be able to practice in the classroom to promote achievement levels as outlined in Florida's A+ Plan. Ten major domains were recognized. Of those 10 domains, two were rated as high priority and dealt approximately with the following areas: instruction (Bandura, 1997; Fullan, 1991; Putnam & Borko, 1997) and classroom-based competencies (Newmann, King, & Youngs, 2000; Wenglinsky, 2002; Zeichner, 1993). These two domains were the principal emphasis of the NTA.

Methodology

There were two research questions that this study intended to answer about the instrument so that results may began to assist in defining it for future generalizations back to the K-12 and college and university teacher training populations:

1) What is the dimensionality structure of the instrument?

2) Does the instrument develop constructs that will measure the theorized domains?

Using a four-point Likert-type scale, where 1 = Not Adequately Prepared; 2 = Somewhat Prepared; 3 = Prepared; and 4 = Very Prepared, the instrument consisted of 16 items, which asked respondents to indicate how prepared they felt to perform various classroom instructional and management tasks (Appendix A).

Reliability

Using the software Analysis of Moment Structures (AMOS) version 4.01 (Arbuckle, 1999), a model was created to obtain measurement reliability estimates based on 2000 bootstrapped samples. The reliability estimates for the instrument's scores were very high, which meant that this instrument had internal consistency and the items on the instrument shared a large percentage of the variance. For the NTA group, the estimated reliability coefficient = .920 with bootstrapped 90% lower and upper confidence limits of (.895, .937). For the Non-NTA group, the estimated reliability coefficient = .922 with bootstrapped 90% lower and upper confidence limits of (.878, .947). The small width found in both bootstrapped confidence limits indicates that there was stability in the sample measurement reliabilities and, thus, estimates based on these samples had a high probability of stability upon replication.

As a medium to allow others to implement further testing of the instrument, or produce competing models, means and standard deviations are provided pertaining to the participants' responses to the 16 items in Table 2. Pearson correlations of the 16 items are presented in Table 3. Because of the number of statistical tests performed, a Bonferroni correction of alpha = .001 was utilized to ensure that the possibility of false rejections was not too great.

Table 2. Descriptive Statistics of Participants' Responses to Questions.

Item	M	SD
1	2.53	.72
2	2.56	.84
3	2.51	.83
4	2.66	.96
5	2.39	.90
6	2.67	.88
7	2.65	.80
8	2.57	.72
9	2.58	.91
10	2.67	.70
11	2.77	.92
12	2.56	.85
13	2.68	.75
14	2.80	.65
15	2.49	.87
16	3.19	.67

The scale needed to be validated to determine if it measured the two domains and if these domains held together. Factor analysis reduces the number of original variables, 16 in this case, into a smaller set of factors to obtain parsimonious dimensionality. Thus, there will be an attempt to capture as much of the variation among the 16 variables as possible with the least amount of dimensions. However, there is a cost and benefit situation to consider. How much loss in precision of the original variables will be tolerated (i.e., the cost) for the benefit of attaining a more parsimonious solution? It was felt that a two dimensional structure would exhibit the nature of the 16 variables, and also that the variance of each variable would be captured sufficiently by the factor structure. That is, all individual variables would be well represented.

A confirmatory factor analysis (CFA), using the extraction method of maximum likelihood with oblimin rotation, was conducted to look at the total variance explained by the model.

Table 3. Correlation Matrix.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	---	.49	.60	.49	.28	.52	.52	.59	.49	.68	.40	.49	.47	.53	.43	.28
2	.49	---	.70	.52	.34	.41	.43	.34	.29	.48	.31	.37	.37	.51	.50	.19
3	.60	.70	---	.65	.46	.37	.46	.43	.35	.52	.50	.38	.46	.47	.56	.23
4	.49	.52	.65	---	.62	.21	.39	.23	.23	.44	.38	.22	.34	.33	.42	.18
5	.28	.34	.46	.62	---	.09	.17	.14	.17	.34	.45	.18	.26	.24	.26	.13
6	.52	.41	.37	.21	.09	---	.57	.59	.54	.60	.22	.66	.53	.56	.42	.41
7	.52	.43	.46	.39	.17	.57	---	.48	.47	.50	.39	.55	.59	.41	.46	.19
8	.59	.34	.43	.23	.14	.59	.48	---	.48	.56	.39	.49	.44	.53	.38	.33
9	.49	.29	.35	.23	.17	.54	.47	.48	---	.56	.42	.76	.67	.56	.37	.22
10	.68	.48	.52	.44	.34	.60	.50	.56	.56	---	.51	.54	.57	.65	.52	.38
11	.40	.31	.50	.38	.45	.22	.39	.39	.42	.51	---	.35	.49	.39	.42	.22
12	.49	.37	.38	.22	.18	.66	.55	.49	.76	.54	.35	---	.70	.52	.41	.30
13	.47	.37	.46	.34	.26	.53	.59	.44	.67	.57	.49	.70	---	.53	.54	.25
14	.53	.51	.47	.33	.24	.56	.41	.53	.56	.65	.39	.52	.53	---	.46	.30
15	.43	.50	.56	.42	.26	.42	.46	.38	.37	.52	.42	.41	.54	.46	---	.43
16	.28	.19	.23	.18	.13	.41	.19	.33	.22	.38	.22	.30	.25	.30	.43	---

Because the scores from the items on the instrument were correlated moderately, it was theorized that the underlying factors for these items were correlated as well. Therefore, oblimin rotation was used, which permits the factors to be correlated and adds to the simplicity and the generalizability of results (Pedhazur & Schmelkin, 1991).

The extent of the correlation between factors was predetermined at ≥ .350 based on the researcher's prediction that the degree of correlation would remain in the moderate to high range. Although high correlations of the ilk ≥ .700 are preferred, the scholarly literature has indicated that loadings between .300 and .500 are often the norm (cf. Biggs, Kember, & Leung, 2001). The variables were reasonably multivariate normal. To determine if it was appropriate to proceed with a factor analysis, an examination of the correlation matrix established that the variables of study were sufficiently related to one another, to a degree significantly different than the identity matrix (Bartlett's Test of Sphericity $\chi^2 = 901.347(120)$; $p < .001$).

In terms of the goodness-of-fit of the model to the sample data, large values of chi

square (χ^2) mean that the model is a bad fit for the data and small values signify that the data is a good fit. The study's sample size of $\underline{n}$ = 105 appears to be ample enough in terms of adhering to the principle of having "... the minimum number of subjects required is 5-10 times the number of observed indicators" (Bryant & Yarnold, 1995, p. 117).

Taking sample size into account, the use of only the χ^2 statistic as a measure of fit may render uncertainty concerning the overall appropriateness of the study's model. Thus, a χ^2 change test was conducted, which compared the values for χ^2 from a one-factor solution, a two-factor solution, and a three-factor solution. Further, the ratio of χ^2 to degrees of freedom (χ^2/df ratio) was used to compare the relative fit of the three models. As the χ^2/df ratio decreases, the fit of a model is improved (Hoelter, 1983).

The one-factor solution had χ^2 = 261.160(104); $p < .001$; 2.51 χ^2/df ratio, the two-factor had χ^2 = 140.558(89); $p < .001$; 1.58 χ^2/df ratio, or a χ^2 change of 120.602, and the three-factor had χ^2 = 98.114(75); $p < .05$; 1.31 χ^2/df ratio, or a χ^2 change of 42.444. The highly statistically significant change test for the two-factor solution indicated that it fit the data better than a one-factor or three-factor solution, where the latter factor solution did not indicate a more significant change by adding a third factor to the model. Also, the χ^2/df ratio was very similar between the two-factor (1.58) and the three-factor (1.31) models. The two-factor model was preferred because of its more simple nature and the fact that the three-factor, more complex model did not appear to offer much more substantial data about model fit.

As advocated by Mulaik et al. (1989) and Tanaka (1993), various indicators of fit were utilized, beyond the χ^2 criterion of fit or no fit, to examine the multiple aspects that may encompass a model and also to determine how closely the model fits the data. Arbuckle's (1999) software AMOS was used to specify the model. As relative fit measures, the incremental fix index (IFI = .977), the comparative fit index (CFI = .977), and the Tucker-Lewis index (TLI = .969) all indicated that the proposed model compared very well to, and exceeded, a null model per the cut point fixed at ≥ .95 (Hu & Bentler, 1999), which was established due to lower magnitudes of a few of the factor loadings. For all fit indices, a rigid cut point was necessary to yield a rejection rate for the few instances where there were low loading circumstances.

For indices based on χ^2, or an absolute fit measure, the root mean square error of approximation (RMSEA) ranges from 0 to 1, with scores of .05, .08, and, .10 representing the magnitude of population misfit (Browne & Cudeck, 1993). This index can also serve as a noncentrality-based fit index. For this model, the RMSEA = .104, meaning that this model was a fairly good estimation of misfit to the population correlation matrix, but did have some error. The expected cross-validation index (ECVI) was 3.049 (90% CI 2.680, 3.492), which is an approximated measure of the goodness-of-fit that the present model would attain in an additional sample of the same size.

To determine how many factors to retain, multiple decision rules were used (Thompson & Daniel, 1996). The traditional eigenvalue greater than 1.00 rule (K1) was analyzed as the lower boundary for the number of factors to be retained (Guttman, 1954; Kaiser, 1960). However, this method of extraction has been noted to both overestimate (Hakstian, Rogers, & Cattell, 1982; Zwick & Velicer, 1986) and underestimate (Cattell & Vogelmann, 1977; Hakstian et al., 1982) the number of factors retained and yield false support for classifying scales as multidimensional (Bernstein & Teng, 1989).

A second method was used with a scree plot (Cattell, 1966). In this technique, the total factors retained were based on the number of eigenvalues that fell before the last major drop on the scree plot. This method potentially could lend itself to subjectivity and poor decisions in terms of the number of factors to retain due to its variability of results and, thus, reliability (Zwick & Velicer, 1986). Yet, results indicated that the scree test produced limited accuracy (Zwick & Velicer, 1982; 1986).

In a third method, a parallel analysis (PA) was run on the data and factors were retained based on a comparison between the scree plot from the random data generated via the PA and the scree plot from the actual data.

Factors from the actual data that had eigenvalues greater than the eigenvalues produced from the PA were extracted because they exceeded chance levels of the eigenvalues from the PA and, thus, indicated that they were "authentic" factors (Horn, 1965; Thompson & Daniel, 1996). This technique has yielded fairly accurate results (Zwick & Velicer, 1986). Finally, Velicer's (1976) Minimum Average Partial (MAP) method was utilized. Using a matrix of partial correlations from the study, the average of the partial squared correlation was determined. When the smallest average squared correlation was attained, no more factors were removed. This extraction method has been found to be very accurate, especially when compared against the traditional K1 rule (Zwick & Velicer, 1982).

Based on the implementation of multiple decision rules and splitting the data in half to determine if the number of factors extracted replicated on all of the multiple decision rules applied, it was determined that two factors should be extracted for the model. The variable (p) to factor (m) ratio was 8:1, where the number of variables was a constant at 16 and the number of factors extracted was 2. This p: m ratio has been cited as reasonable for practical usage (Zwick & Velicier, 1986). The variance of the first factor was = 7.531 and the second factor = 1.789. The two eigenvalues had a cumulative percentage = 58.247. They accounted for 58% of the variation among the 16 variables. The correlation between factor 1 and factor 2 was .526.

To name these two factors, the solution was rotated to simulate a simple structure via oblimin rotation. This will yield the relative contribution of each variable to a factor by correlating variables to factors. The pattern coefficients are standardized regression weights that account for the correlation among the two factors and the structure coefficients are bivariate correlations between the two factors and the 16 variables.

Examination of both the pattern (p) and structure (s) coefficients rendered like interpretations of the factor structure. In terms of convergent validity, how a factor primarily influenced a variable was established as both $p \geq .700$ and $s = \geq .700$, while a more moderate extent influence was established as both p and s between .350 and .699. Factor 1 appeared to influence principally X_6 ($p_6 = .807$; $s_6 = .762$), X_9 ($p_9 = .891$; $s_9 = .800$), X_{12} ($p_{12} = .933$; $s_{12} = .839$), and X_{13} ($p_{13} = .757$; $s_{13} = .780$). It influenced to a moderate degree X_1 ($p_1 = .485$; $s_1 = .679$), X_7 ($p_7 = .586$; $s_7 = .671$), X_8 ($p_8 = .630$; $s_8 = .667$), X_{10} ($p_{10} = .615$; $s_{10} = .758$), X_{14} ($p_{14} = .614$; $s_{14} = .706$), X_{16} ($p_{16} = .459$; $s_{16} = .448$), and X_{15} ($p_{15} = .375$; $s_{15} = .579$). Factor 1 had a lesser influence on X_{11} ($p_{11} = .302$; $s_{11} = .503$). Both X_{11} and X_{15} were shared with Factor 2.

Due to this result, Factor 1 should be named Classroom and Behavior Management. This incorporated in-class activities, which addressed issues that impacted both learning and instruction such as motivating students to behave, implementing techniques to accommodate various learning styles, and promoting an effective learning environment. This combination of subject matter and pedagogical knowledge has been found to enable teachers to understand and explain content-related tasks and concepts connected to student learning (Beijaard, 1995; Bennett & Carre, 1993).

Factor 2 seemed to influence primarily X_3 ($p_3 = .768$; $s_3 = .849$) and X_4 ($p_4 = .814$; $s_4 = .785$). To a moderate degree, it influenced X_2 ($p_2 = .617$; $s_2 = .714$), X_5 ($p_5 = .646$; $s_5 = .600$), X_{11} ($p_{11} = .384$; $s_{11} = .542$), and X_{15} ($p_{15} = .389$; $s_{15} = .586$), with both of the latter two variables shared with Factor 1. Factor 2 should be named Instructional Knowledge and Skills, which looked at questions that measured if teachers thought they were prepared to teach students the content standards deemed important toward achieving grade level proficiency. Teacher preparedness in terms of content knowledge has been found to inform classroom learning, which affects instructional decisions (Swafford, Chapman, Rhodes, & Kallis, 1996).

In general, there was a rotation that separated the variables in a manner in which highly correlated variables had sufficient factor pattern coefficients on one factor and very little on a second factor, or discriminant validity was established. In fact, only two variables, X_{11} and X_{15}, had factor pattern coefficients split on more than one factor.

This is important for future theoretical use and measurement of the scale, where dimension one separated classroom and behavior management items from dimension two related to instructional knowledge items.

The two dimensional structure appeared to capture the 16 variables. Now, however, were there individual variables that were not well represented in the structure? Communalities (h^2) are the proportion of each variable explained by the factor structure (i.e., akin to R^2). Extraction communalities ranged from .201 to .721. For example, X_3 had the highest $h^2 = .721$. This is the percentage of variation of this variable that is accounted for by the factor solution. X_{16} had the lowest communality at .201. If the cut point of $h^2 \geq .350$ is used, which was previously implemented in the study, to examine these communalities, all of the variables, with the exception of X_{16}, were accounted for noticeably by the factor solution.

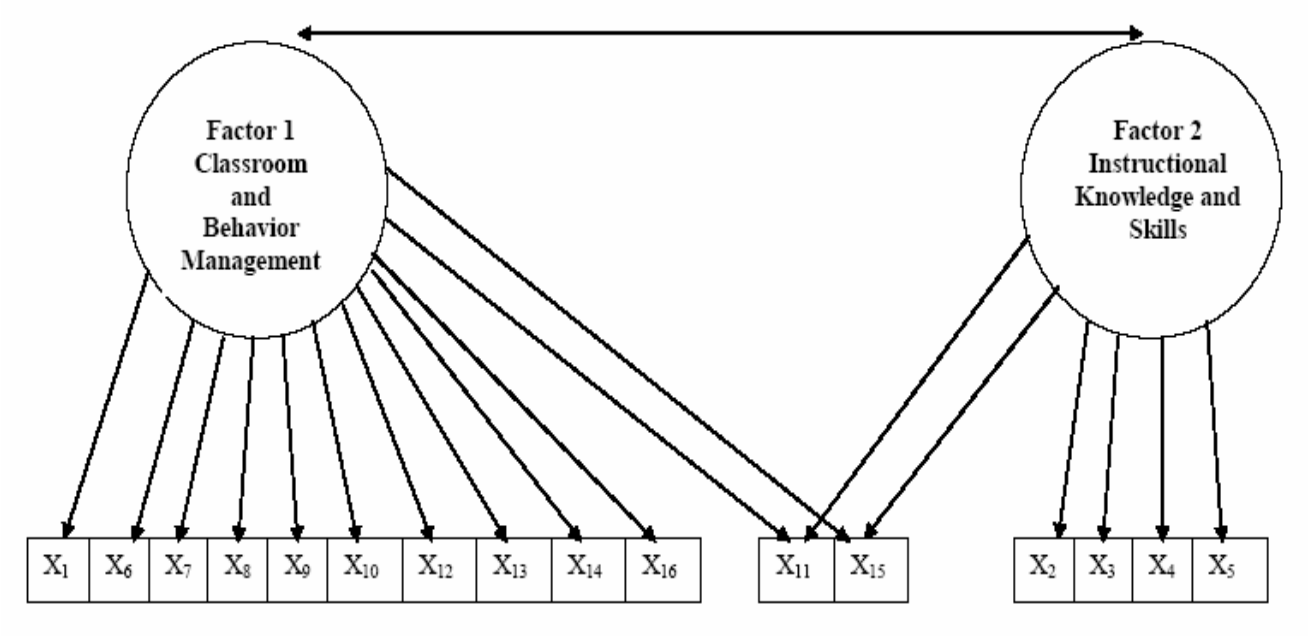
Looking at X_{16}, its unique variance was .799 (i.e., 1 - .201), which indicated that 80% of this variable's variance was unexplained by factor one. However, this variable's pattern and structure coefficients were acceptable, signifying that X_{16}'s factor had a moderate influence on it, but was less sufficient at predicting the amount of variance pertaining to the variable. Yet, given the high score reliability of the instrument at .920 and .922 for both groups, there appears to be little error and, therefore, the large unique variance for X_{16} should not be attributed extensively to measurement error.

Conclusion

In terms of the sample size, and admittedly a border-line size, there were a number of techniques previously-mentioned (e.g., χ^2 change test, various indicators of fit were utilized beyond the χ^2 criterion, splitting the data in half to determine if the number of factors extracted replicated on all of the multiple decision rules applied, etc.) throughout the study to monitor size to establish if it had a substantial influence on the results. It appeared that this study's sample size was within a suitable range of the number of subjects per observed indicators.

The findings of this research suggest that the NTA scale was measured as a multidimensional instrument with two distinct factors. This implied that one factor was not adequate for the entire instrument. A CFA corroborated that the instrument had construct validity by providing evidence that these two domains held together and had a set of 16 items that were relatively homogeneous. These findings assisted in answering the study's two research questions: what is the dimensionality structure of the instrument and does the instrument develop constructs that will measure the theorized domains? The preliminary findings connected to these questions are salient because they suggest that this instrument has an adept developmental foundation both in terms of measurement and substance. To be sure, more validation of scores needs to be secured across many implementations of this instrument, but early development appears promising.

Figure 1: Two-Dimensional Model



References

Arbuckle, J. L. (1999). *AMOS* (Version 4.01). Chicago: SmallWaters Corporation.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Beijaard, D. (1995). Teachers' prior experiences and actual perceptions of professional identity. *Teachers and Teaching: Theory and Practice, 1*, 281-294.

Bennett, N., & Carre, J. C. (1993). *Learning to teach*. London: Routledge.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Biggs, J., Kember, D., & Leung, D. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71,* 133-149.

Broward County Public Schools. (2003). *Interesting factoids*. Retrieved January 7, 2003, from http://www.browardschools.com/about /factoids.htm

Brown, B., & Campion, M. A. (1994). Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology*, *79*, 897-908.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models*, p. 136-162. Thousand Oaks, CA: Sage.

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics*, p. 99-136. Washington, D.C.: American Psychological Association.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12*, 289-325.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*, 947-967.

Florida Atlantic University. (2002). *Questions and answers about CLAST*. http://www.fau.edu/student/student/testing/clast_page2.html

Florida Department of Education. (2000). *CLAST technical report: 1999-2000*. Tallahassee, FL: State of Florida Department of State.

Florida International University. (2002). *CLAST*. Retrieved November 24, 2002, from http://www.fiu.edu/~testing/clast.htm

Fullan, M. G. (1991). *The new meaning of educational change*. New York: Teachers College Press.

Fullan, M. G. (2000). The return of large-scale reform. *Journal of Educational Change,1*, 5-28.

Garet, M. S., Porter, A. G., Desimone, L., Birman, B. F., & Yoon, K. S (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research, 17*, 193-219.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness of fit indices. *Sociological Methods and Research, 11*, 325-344.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: John Wiley & Sons, Inc.

Huberty, C. J., & Lowman, L. L. (1998). *Discriminant analysis in higher education research*. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 181-234). New York: Agathon Press.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60*, 543-563.

Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. *Multivariate Behavioral Research, 22*, 307-329.

Indian River Community College. (2002). *CLAST*. Retrieved November 24, 2002, from http://www.ircc.cc.fl.us/educserv/admis/tests/clast.html

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics, 23*, 639-645.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*, 1-11.

Lieberman, M. G., & Morris, J. D. (2004, April). *Selecting predictor variables in logistic regression*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Meehl, P. E. (1993). Four queries about factor reality. *History and Philosophy of Psychology Bulletin, 5*, 4-5.

Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement, 55*, 438-441.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S, & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Newmann, F. M., King, M. B., & Youngs, P. (2000, April). *Professional development that addresses school capacity: Lessons from urban elementary schools*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*, 699-705.

Putnam, R., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *The international handbook of teachers and teaching*, p. 1223-1296. Dordrecht, The Netherlands: Kluwer.

Roth, P. L., & Bobko, P. (2000). College grade point average as a personnel selection device: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology, 85*, 399-406.

Swafford, J., Chapman, V., Rhodes, R., & Kullis, M. (1996). A literate analysis of trends in literacy education. In D. J. Leu, C. K. Kinzer, & K. A. Hinchman (Eds.), *Literacies for the 21st century: Research and practice*, p. 437-446. Chicago: National Reading Conference.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In J. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*, p. 10-39. Newbury Park, CA: Sage.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

University of South Florida. (2002). *CLAST*. Retrieved November 24, 2002, from http://usfweb.usf.edu/ugrads/eandt/clast_state_document.htm

Velicer, W. F. (1976). The relation between factor score estimates, image scores and principal component scores. *Educational and Psychological Measurement, 36*, 149-159.

Wenglinsky, H. (2002, February 13). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(2). http://epaa.asu.edu/epaa/v10n12

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*, 177-185.

Zeichner, K. M. (1993). Traditions of practice in U.S. preservice teacher education programs. *Teaching and Teacher Education, 9*, 1-13.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. M*ultivariate Behavioral Research, 17*, 253-269.

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Appendix A

X_1: Identify how individual differences and learning styles affect instructional delivery.
X_2: Recognize Grade Level Expectations (GLE).
X_3: Recognize Critical Content (CC).
X_4: Recognize Sunshine State Standards (SSS).
X_5: Recognize the Florida Comprehensive Test (FCAT).
X_6: Develop strategies to motivate students to learn.
X_7: Advance the delivery of instruction through effective organization and time management skills.
X_8: Identify effective teaching behaviors.
X_9: Develop strategies to diminish misbehavior.
X_{10}: Identify individual differences and learning styles.
X_{11}: Develop effective record keeping routines.
X_{12}: Acquire strategies to motivate students to behave.
X_{13}: Promote positive classroom behavior through effective organization and time management skills.
X_{14}: Demonstrate teaching and learning behaviors that promote an effective learning environment.
X_{15}: Develop goals that are realistic and achievable for your Professional Growth Plan (PGP).
X_{16}. *Work* cooperatively with students, colleagues, administrators, and parents.